
An Information Criterion for Inferring Coupling of Distributed Dynamical Systems

Oliver M. Cliff*

Mikhail Prokopenko[†]

Robert Fitch^{*,‡}

*Australian Centre for Field Robotics,
The University of Sydney, Australia,
o.cliff@acfr.usyd.edu.au

[†]Complex Systems Research Group,
The University of Sydney, Australia,
mikhail.prokopenko@sydney.edu.au

[‡]Centre for Autonomous Systems,
University of Technology Sydney, Australia,
robert.fitch@uts.edu.au

Abstract

The behaviour of many real-world phenomena can be modelled by nonlinear dynamical systems whereby a latent system state is observed through a filter. We are interested in interacting subsystems of this form, which we model by a set of coupled maps as a synchronous update graph dynamical system. Specifically, we study the structure learning problem for spatially distributed dynamical systems coupled via a directed acyclic graph. Unlike established structure learning procedures that find locally maximum posterior probabilities of a network structure containing latent variables, our work exploits the properties of dynamical systems to compute globally optimal approximations of these distributions. We arrive at this result by the use of time delay embedding theorems. Taking an information-theoretic perspective, we show that the log-likelihood has an intuitive interpretation in terms of information transfer.

1 Introduction

Complex systems are broadly defined as systems that comprise interacting nonlinear components [1]. Discrete-time complex systems can be represented using graphical models such as *graph dynamical systems (GDSs)* [2, 3], where spatially distributed dynamical units are coupled via a directed graph. The task of learning the structure of such a system is to infer directed relationships between variables; in the case of dynamical systems, these variables are typically hidden [4]. In this paper, we study the *structure learning* problem for complex networks of nonlinear dynamical systems coupled via a directed acyclic graph (DAG). Specifically, we formulate synchronous update GDSs as dynamic Bayesian networks (DBNs) and study this problem from the perspective of information theory.

The structure learning problem for distributed dynamical systems is a precursor to inference in systems that are not fully observable. This case encompasses many practical problems of known artificial, biological and chemical systems, such as neural networks [5, 6, 7], multi-agent systems [8, 9, 10, 11] and various others [1]. Modelling a partially observable system as a dynamical network presents a challenge in synthesising these models and capturing their global properties [1]. In addressing this challenge, we draw on probabilistic graphical models (specifically Bayesian network (BN) structure learning) and nonlinear time series analysis (differential topology).

In this paper we exploit the properties of discrete-time multivariate dynamical systems in inferring coupling between latent variables in a DAG. Specifically, the main focus of this paper is to analytically derive a measure (score) for evaluating the fitness of a candidate DAG, given data. We assume the

data are generated by a certain family of multivariate dynamical system and are thus able to overcome the issue of latent variables faced by established structure learning algorithms. That is, under certain assumptions of the dynamical system, we are able to employ time delay embedding theorems [12, 13] to compute our scores.

Our main result is a tractable form of the log-likelihood function for synchronous GDSs. Using this result, we are able to directly compute the *Bayesian information criterion (BIC)* [14] and *Akaike information criterion (AIC)* [15] and thus achieve globally optimal approximations of the posterior distribution of the graph. Finally, we show that the log-likelihood and log-likelihood ratio can be expressed in terms of *collective transfer entropy* [16, 5]. This result places our work in the context of effective network analysis [17, 18] based on information transfer [19, 6, 20, 10] and relates to the information processing intrinsic to distributed computation [21].

2 Related Work

We are interested in classes of systems whereby dynamical units are coupled via a graph structure. These types of systems have been studied under several names, including complex dynamical networks [1], spatially distributed dynamical systems [4, 7], master-slave configurations (or systems with a skew product structure) [22], and coupled maps [23]. Common to each of these definitions is that the multivariate state of the system comprises individual subsystem states, the dynamics of which are given by a set of either discrete-time maps or first-order ordinary differential equations (ODEs), called a flow. We assume the discrete-time formulation, where a map can be obtained numerically by integrating differential equations or recording experimental data (observations) at discrete-time intervals [4]. The literature on coupled dynamical systems is often focused on the analysis of characteristics such as stability and synchrony of the system. In this work we draw on the fields of BN structure learning and nonlinear time series analysis to infer coupling between spatially distributed dynamical systems.

BN structure learning comprises two subproblems: *evaluating* the fitness of a graph, and *identifying* the optimal graph given this fitness criterion [24]. The evaluation problem is particularly challenging in the case of graph dynamical systems, which include both latent and observed variables. A number of theoretically optimal techniques exist for the evaluation problem for BNs with complete data [25, 26, 27], which have been extended to DBNs [28]. With incomplete data, however, the common approach is to resort to approximations that find local optima, e.g., expectation-maximisation (EM) [28, 29]. An additional caveat with respect to structure learning is that algorithms find an equivalence class of networks with the same Markov structure, and not a unique solution [24].

In nonlinear time series analysis, the problem of inferring coupling strength and causality in complex systems has received significant attention recently [30, 31]. Early work by Granger defined causality in terms of the predictability of one system linearly coupled to another [32]. Although this measure is popular for identifying coupling, it requires systems are linear statistical models and is considered insufficient for inferring coupling between dynamical systems due to inseparability [33]. Another method popular in neuroscience is transfer entropy, which was introduced to quantify the information transfer between nonlinear (finite-order Markov) systems [30]. Transfer entropy has been used to recover interaction networks in numerous fields such as multi-agent systems [10] and effective networks in neuroscience [5, 6, 34]. More recently, researchers have used the additive noise model [31, 35] to infer unidirectional cause and effect relationships with observed random variables and find a unique DAG (as opposed to an equivalence class). These studies have been extended by exploring weakly additive noise models for learning the structure of systems of observed variables with nonlinear coupling [36].

A recent approach to inferring causality is convergent cross-mapping (CCM), which is based on Takens theorem [37] and tests for causation (predictability) by considering the history of observed data of a hidden variable in predicting the outcome of another [33]. Using a similar approach, Schumacher et al. [7] used Stark’s bundle delay embedding theorem [38, 12] to predict one subsystem from another using Gaussian processes. This algorithm can thus be used to infer the driving systems in spatially distributed dynamical systems in a similar manner to our work. However, both papers do not consider the problem of inference over the entire network structure, or formally derive the measures used therein. In our work, we provide a rigorous proof based on established structure learning procedures and discuss the problem of inference within a distributed dynamical system.

3 Background

This section summarises relevant technical concepts used throughout the paper. First, a stochastic temporal process X is defined as a sequence of random variables (X_1, X_2, \dots, X_N) with a realisation (x_1, x_2, \dots, x_N) for countable time indices $n \in \mathbb{N}$. Consider a collection of M processes, and denote the i th process X^i to have associated realisation x_n^i at temporal index n , and \mathbf{x}_n as all realisations at that index $\mathbf{x}_n = \langle x_n^1, x_n^2, \dots, x_n^M \rangle$. If X_n^i is a discrete random variable, the number of values the variable can take on is denoted $|X_n^i|$. The following sections collect results from DBN literature, attractor reconstruction, and information theory that are relevant to this work.

3.1 Dynamic Bayesian networks (DBNs)

DBNs are a general graphical representation of a temporal model, representing a probability distribution over infinite trajectories of random variables $(\mathbf{Z}_1, \mathbf{Z}_2, \dots)$ compactly. These models are a more expressive framework than the hidden Markov Model and Kalman filter model (or linear dynamical system) [28]. In this work, we denote $\mathbf{Z}_n = \{\mathbf{X}_n, \mathbf{Y}_n\}$ as the set of hidden and observed variables, respectively, where $n \in \{1, 2, \dots\}$ is the temporal index.

BNs $B = (G, \Theta)$ represent a joint distribution $p(\mathbf{z})$ graphically and consist of: a DAG G and a set of conditional probability distribution (CPD) parameters Θ . DBNs $B = (B_1, B_{\rightarrow})$ extend the BN to model temporal processes and comprise two parts: the prior BN $B_1 = (G_1, \Theta_1)$, which defines the joint distribution $p_{B_1}(\mathbf{z}_1)$; and the *two-time-slice Bayesian network (2TBN)* $B_{\rightarrow} = (G_{\rightarrow}, \Theta_{\rightarrow})$, which defines a first-order Markov process $p_{B_{\rightarrow}}(\mathbf{z}_{n+1} | \mathbf{z}_n)$ [28]. This formulation allows for a variable to be conditioned on its respective parent set $\Pi_{G_{\rightarrow}}(\mathbf{Z}_{n+1}^i)$ that can come from the preceding time slice or the current time slice, as long as G_{\rightarrow} forms a DAG. The 2TBN probability distribution factorises according to G_{\rightarrow} with a local CPD p_D estimated from an observed dataset. That is, given a set of stochastic processes $(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_N)$, the realisation of which constitutes a dataset $D = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N)$, we obtain the 2TBN distribution as

$$p_{B_{\rightarrow}}(\mathbf{z}_{n+1} | \mathbf{z}_n) = \prod_i p_{B_{\rightarrow}}(z_{n+1}^i | \pi_{G_{\rightarrow}}(\mathbf{Z}_{n+1}^i)), \quad (1)$$

where $\pi_{G_{\rightarrow}}(\mathbf{Z}_{n+1}^i)$ denotes the (index-ordered) set of realisations $\{z_o^j : Z_o^j \in \Pi_{G_{\rightarrow}}(\mathbf{Z}_{n+1}^i)\}$.

3.2 Embedding theory

Embedding theory refers to methods from differential topology for inferring the (hidden) state of a dynamical system from a reconstructed sequence of observations. The state of a discrete-time dynamical system is given by a point \mathbf{x}_n confined to a d -dimensional manifold \mathcal{M} . The time evolution of this state is described by a map $f : \mathcal{M} \rightarrow \mathcal{M}$, so that the sequence of states (\mathbf{x}_n) is given by $\mathbf{x}_{n+1} = f(\mathbf{x}_n)$. In many situations we only have access to a filtered, scalar representation of the state, i.e., the measurement $y_n = \psi(\mathbf{x}_n)$ given by some *measurement function* $\psi : \mathcal{M} \rightarrow \mathbb{R}$ [37, 38].

The celebrated Takens' theorem [37] shows that for typical f and ψ , it is possible to reconstruct f from the observed time series up to some smooth coordinate change. More precisely, fix some κ (the *embedding dimension*) and some τ (the *time delay*), then define the *delay embedding map* $\Phi_{f,\psi} : \mathcal{M} \rightarrow \mathbb{R}^\kappa$ by

$$\Phi_{f,\psi}(\mathbf{x}_n) = y_n^{(\kappa)} = \langle y_n, y_{n-\tau}, y_{n-2\tau}, \dots, y_{n-(\kappa-1)\tau} \rangle. \quad (2)$$

In differential topology, an *embedding* refers to a smooth map $\Psi : \mathcal{M} \rightarrow \mathcal{N}$ between manifolds \mathcal{M} and \mathcal{N} if it maps \mathcal{M} diffeomorphically onto its image; therefore, $\Phi_{f,\psi}$ has a smooth inverse $\Phi_{f,\psi}^{-1}$. The implication of Takens' theorem is that for typical f and ψ , the image $\Phi_{f,\psi}(\mathcal{M})$ of \mathcal{M} is completely equivalent to \mathcal{M} itself, apart from the smooth invertible change of coordinates given by the mapping $\Phi_{f,\psi}$. An important consequence of this theorem is that we can define a map $\mathbf{F} = \Phi_{f,\psi} \circ f \circ \Phi_{f,\psi}^{-1}$ on $\Phi_{f,\psi}$, such that $y_{n+1}^{(\kappa)} = \mathbf{F}(y_n^{(\kappa)})$ [38].

There are technical assumptions for Takens' theorem (and the generalised versions employed herein) to hold. These assumptions require: (f, ψ) to be generic functions (in terms of Baire space), a restricted number of periodic points, and distinct eigenvalues at each neighbourhood of these points [37, 38, 12, 13].

3.3 Information theoretic measures

Conditional entropy represents the uncertainty of a random variable X after taking into account the outcomes of another random variable Y by

$$H(X | Y) = - \sum_{x,y} p(x,y) \log_2 p(x | y). \quad (3)$$

Multivariate *transfer entropy* is a measure that computes the information transfer from a set of source processes to a set of destination process [6]. In this work, we use the formulation of collective transfer entropy [16], where the information transfer from m source processes $\mathbf{V} = \{Y^1, Y^2, \dots, Y^m\}$ to a single destination process Y can be decomposed as a sum of conditional entropy terms:

$$T_{\mathbf{V} \rightarrow Y} = H(Y_{n+1} | Y_n^{(\kappa)}) - H(Y_{n+1} | Y_n^{(\kappa)}, \langle Y_n^{i, (\kappa^i)} \rangle), \quad (4)$$

where $Y_n^{i, (\kappa^i)} = \langle Y_n^i, Y_{n-\tau^i}^i, Y_{n-2\tau^i}^i, \dots, Y_{n-(\kappa^i-1)\tau^i}^i \rangle$ for some κ^i and τ^i , and similarly for $Y_n^{(\kappa)}$.

4 Representing nonlinear dynamical networks as DBNs

We express multivariate dynamical systems as a synchronous update GDS to allow for generic maps. With this model, we can express the time evolution of the GDS as a stationary DBN, and perform inference and learning on the subsequent graph. We formally state the network of dynamical systems as a special case of the sequential GDS [39] with an observation function for each vertex.

Definition 1 (Synchronous graph dynamical system (GDS)). *A synchronous GDS is a tuple $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$ that consists of:*

- a finite, directed graph $G = (\mathcal{V}, \mathcal{E})$ with edge-set $\mathcal{E} = \{E^i\}$ and M vertices comprising the vertex set $\mathcal{V} = \{V^i\}$;
- a multivariate state $\mathbf{x}_n = \langle x_n^i \rangle$, composed of states for each vertex V^i confined to a d^i -dimensional manifold $x_n^i \in \mathcal{M}^i$;
- an M -variate observation $\mathbf{y}_n = \langle y_n^i \rangle$, composed of scalar observations for each vertex $y_n^i \in \mathbb{R}$;
- a set of local maps $\{f^i\}$ of the form $f^i : \mathcal{M} \rightarrow \mathcal{M}^i$, which update synchronously and induce a global map $f : \mathcal{M} \rightarrow \mathcal{M}$; and
- a set of local observation functions $\{\psi^1, \psi^2, \dots, \psi^M\}$ of the form $\psi^i : \mathcal{M}^i \rightarrow \mathbb{R}$.

Without loss of generality, we can use local functions to describe the time evolution of the subsystems:

$$x_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j) + v_{f^i} \quad (5)$$

$$y_{n+1}^i = \psi^i(x_{n+1}^i) + v_{\psi^i}. \quad (6)$$

Here, v_{f^i} is i.i.d. additive noise and v_{ψ^i} is noise that is either i.i.d. or dependent on the state, i.e., $v_{\psi^i}(x_{n+1}^i)$. The subsystem dynamics (5) are therefore a function of the subsystem state x_n^i and the subsystem parents' state $\langle x_n^{ij} \rangle_j$ at the previous time index such that $f^i : \mathcal{M}^i \times_j \mathcal{M}^{ij} \rightarrow \mathcal{M}^i$. Each subsystem observation is given by (6). We assume the functions $\{f^i\}$ and $\{\psi^i\}$ are invariant w.r.t. time and thus the graph G is stationary.

The time evolution of a synchronous GDS can be modelled as a DBN. First, each subsystem vertex $V^i = \{X_n^i, Y_n^i\}$ has an associated state variable X_n^i and observation variable Y_n^i ; the parents of subsystem V^i are denoted $\Pi_G(V^i)$. Since the graph G_{\rightarrow} is stationary and synchronous, parents of X_{n+1}^i come strictly from the preceding time slice, and additionally $\Pi_{G_{\rightarrow}}(Y_{n+1}^i) = X_{n+1}^i$. Thus, we can build the edge set $\mathcal{E} = \{E^1, E^2, \dots, E^M\}$ in the GDS by means of the DBN. That is, each edge subset E^i is built by the DBN edges

$$E^i = \{V^j \rightarrow V^i : X_n^j \in \Pi_{G_{\rightarrow}}(X_{n+1}^i) \wedge V^j \in \mathcal{V} \setminus V^i\},$$

so long as G forms a DAG.

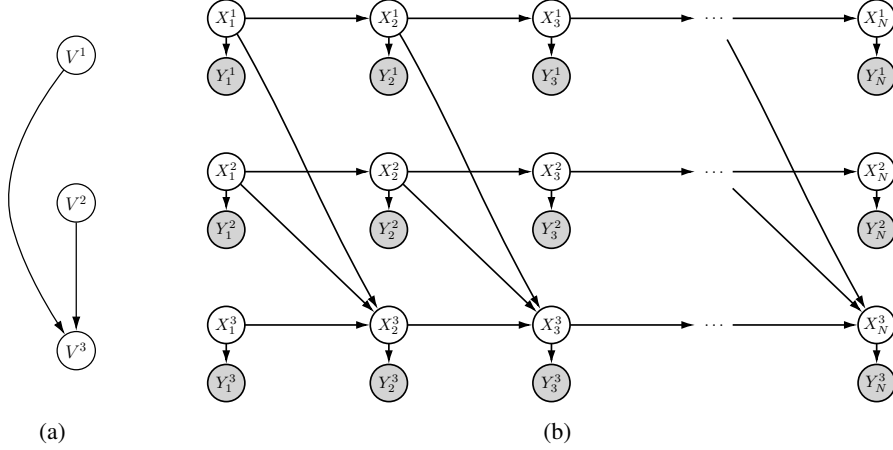


Figure 1: Representation of (a) the GDS with three vertices (V^1 , V^2 and V^3), and (b) the rolled-out DBN of the equivalent structure. Subsystem V^3 is coupled to both subsystems V^1 and V^2 by means of the edges between latent variables $X^1_n \rightarrow X^3_{n+1}$ and $X^2_n \rightarrow X^3_{n+1}$.

As an example, consider the synchronous GDS in Fig. 1(a). The subsystem V^3 is coupled to both subsystem V^1 and V^2 through the edge set $\mathcal{E} = \{V^1 \rightarrow V^3, V^2 \rightarrow V^3\}$. The time-evolution of this network is shown in Fig. 1(b), where the top two rows (processes X^1 and Y^1) are associated with subsystem V^1 , and similarly for V^2 and V^3 . The distributions for the state (5) and observation (6) of M arbitrary subsystems can therefore be factorised according to (1):

$$p_{B_{\rightarrow}}(\mathbf{z}_{n+1} \mid \mathbf{z}_n) = \prod_{i=1}^M p_D(x^i_{n+1} \mid x^i_n, \langle x^{ij}_n \rangle_j) \cdot p_D(y^i_{n+1} \mid x^i_{n+1}). \quad (7)$$

In the rest of the paper we use simplified notation, given this constrained graph structure. Firstly, since our focus is on learning coupling between distributed systems, the superscripts refer to individual *subsystems*, not variables. Thus, although the 2TBN B_{\rightarrow} is constrained such that $\Pi_{G_{\rightarrow}}(Y^i_n) = X^i_n$, the notation Y^{ij}_n denotes the *measurement variable* of the j th parent of subsystem i , e.g., in Fig. 1 an arbitrary ordering of the parents gives $Y^{3,1}_n = Y^1_n$ and $Y^{3,2}_n = Y^2_n$. Secondly, the scoring functions for the 2TBN network B_{\rightarrow} can be computed independently of the prior network B_1 [28]. We will assume the prior network is given, and focus on learning the 2TBN. As a result, we drop the subscript and note that all references to the network B are to the 2TBN. Since B_{\rightarrow} is stationary, learning B_{\rightarrow} is equivalent to learning the synchronous GDS.

5 Learning synchronous GDSs from data

In this section we develop the theory for learning the synchronous update GDS from data. We will focus on techniques for learning graphical models using the *score and search* paradigm, the objective of which is to find a DAG G^* that maximises a score $g(B : D)$. Given such a score, we can then employ established search procedures to find the optimal graph G^* . Thus, we can state that our main goal is to derive a tractable scoring function $g(B : D)$ for synchronous GDSs that gives a parsimonious model for describing the data.

To derive the score we use the DBN formulation of synchronous GDSs (Sec. 4) to show that we cannot directly compute the posterior probability of the network structure (Sec. 5.1). By making some assumptions about the system, however, we are able to compute scores for GDSs by use of attractor reconstruction methods (Sec. 5.2). We conclude this section by giving an interpretation of the log-likelihood in terms of information transfer (Sec. 5.3).

5.1 Structure learning for DBNs

Ideally, we want to be able to compute the posterior probability of the network structure G , given data D . Using Bayes' rule, we can express this distribution as $p(G | D) \propto p(D | G)p(G)$, where $p(G)$ encodes any prior assumptions we want to make about the network G . Thus, the problem becomes that of computing the likelihood of the data, given the model, $p(D | G)$. The likelihood can be written in terms of distributions over network parameters [28]:

$$p(D | G) = \int p(D | G, \Theta)p(\Theta | G)d\Theta, \quad (8)$$

where we denote $\ell(\hat{\Theta}_G : D) = \log p(D | G, \hat{\Theta}_G)$ as the log-likelihood function for a choice of parameters $\hat{\Theta}_G$ that maximise $p(D | G, \Theta)$, given a graph G .

A common approach to compute (8) in closed form is by using Dirichlet priors. This leads to the BD (Bayesian-Dirichlet) score and variants [27, 28]. However, to obtain this analytic solution, we require counts of the tuples $(z_n^i, \pi_G(Z_n^i))$, which involve hidden variables. We will instead use Schwarz's [14] asymptotic approximation of the posterior distribution, which states that

$$\lim_{N \rightarrow \infty} \log p(D | G) \approx \ell(\hat{\Theta}_G : D) - \frac{\log N}{2} C(G) + \mathcal{O}(1), \quad (9)$$

where $C(G)$ is the model dimension (i.e., number of parameters needed for the graph G [28]) and $\mathcal{O}(1)$ is a constant bounded by the number of potential models. The approximation of the posterior (9) requires that data come from an exponential family of likelihood functions with conjugate priors over the model G , and the parameters given the model Θ_G [14].

Akaike gives a similar criterion by approximating the KL-divergence of any model from the data [15]. We can compute both criteria in terms of the log-likelihood function $\ell(\hat{\Theta}_G : D)$ and the model dimension $C(G)$, and thus the problem can be generalised to that of deriving an information criterion for scoring the graph of the form

$$g(B : D) = \ell(\hat{\Theta}_G : D) - f(N) \cdot C(G). \quad (10)$$

When $f(N) = 1$, we have the AIC score [15]; $f(N) = \log(N)/2$ yields the BIC score [14], and $f(N) = 0$ gives the maximum likelihood score.

5.2 Deriving the scores for synchronous GDSs

To calculate the information criterion (10), we require tractable expressions for the log-likelihood function $\ell(\hat{\Theta}_G : D)$ and the model dimension $C(G)$. The form of the CPD in (7) specifies these functions, and for (9) to hold, this distribution must come from an exponential family [14]. We do not assume the underlying model is linear-Gaussian or other known distributions, and thus express the log-likelihood as the maximum likelihood estimate for multinomial distributions [28]. From (7) the log-likelihood then decomposes as

$$\begin{aligned} \ell(\hat{\Theta}_G : D) = -N \sum_{i=1}^M \left[\sum_{x_{n+1}^i} \sum_{\langle x_n^{ij} \rangle_j} p_D(x_{n+1}^i, x_n^i, \langle x_n^{ij} \rangle_j) \log p_D(x_{n+1}^i | x_n^i, \langle x_n^{ij} \rangle_j) \right. \\ \left. + \sum_{x_{n+1}^i} \sum_{y_{n+1}^i} p_D(y_{n+1}^i, x_{n+1}^i) \log p_D(y_{n+1}^i | x_{n+1}^i) \right]. \quad (11) \end{aligned}$$

Note that although we describe the states and observations as discrete in (11), we assume the data are generated by a continuous and stationary process. In theory it is conceivable to have access to an infinite dataset containing realisations of all potential states and observations. In practice we have a limited dataset and therefore must implement a discretisation scheme. Modelling the dynamical systems with non-parametric techniques requires that the number of parameters scales linearly in the size of the data, and thus $C(G)$ scales linearly with N . Instead, later we will assume the observation data are discretised, such that there are $|Y_n^i|$ possible outcomes for an observed random variable Y_n^i .

The log-likelihood function (11) involves distributions over latent variables, and thus we resort to state-space (attractor) reconstruction. First, Lemma 1 shows that a future observation from a given

subsystem can be predicted from a sequence of past observations. Building on this result, we present a computable formulation of the 2TBN distribution $p_{B \rightarrow}(\mathbf{z}_{n+1} \mid \mathbf{z}_n)$ via Lemma 2. We then derive a tractable form of the log-likelihood function, presented in Lemma 1. It is then shown in Theorem 2 that these lemmas allow us to compute the information criterion (10).

Lemma 1. *Consider a synchronous GDS $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$, where the graph G is a DAG. Each subsystem state follows the dynamics $x_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j)$ and emits an observation $y_{n+1}^i = \psi^i(x_{n+1}^i)$; the subsystem observation can be estimated, for some map \mathbf{G}^i , by*

$$y_{n+1}^i = \mathbf{G}^i \left(y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j \right). \quad (12)$$

Proof. Consider a forced system $\mathbf{x}_{n+1} = f(\mathbf{x}_n, \mathbf{w}_n)$ with forcing dynamics $\mathbf{w}_{n+1} = h(\mathbf{w}_n)$ and observation $y_n = \psi(\mathbf{x}_{n+1})$. Given this type of forced system, the bundle delay embedding theorem [38, 12] states that the delay map $\Phi_{f,h,\psi}(\mathbf{x}_n, \mathbf{w}_n) = y_n^{(\kappa)}$ is an embedding for generic f , ψ , and h . Stark [38] proved this result in the case of forcing dynamics h that are independent of the state x .¹ For notational simplicity, we omit dependence on the noise process for the map $\Phi_{f,h,\psi}$; the noise can be considered an additional forcing system so long as v_f is i.i.d and v_ψ is either i.i.d or dependent on the state [12].

Given a DAG G , any ancestor of the subsystem V^i is not dependent on V^i . As such, the sequence

$$y_n^{i,(\kappa^i)} = \Phi_{f^i, \langle f^{ij} \rangle_j, \psi^i} (x_n^i, \langle x_n^{ij} \rangle_j) \quad (13)$$

is an embedding, since $\langle x_n^{ij} \rangle_j$ is independent of x_n^i . Let $\langle x_n^{ijk} \rangle_k$ be the index ordered set of parents of node X_n^{ij} (which itself is the j th parent of the node X_n^i). Under the constraint that G is a DAG, where the state $x_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j) + v_{f^i}$, it follows from the bundle delay embedding theorem [38, 12] that there exists a map \mathbf{F}^i that is well defined and a diffeomorphism between observation sequences. From (13) we can write this map

$$\begin{aligned} y_{n+1}^{i,(\kappa^i)} &= \Phi_{f^i, \langle f^{ij} \rangle_j, \psi^i} \left(f^i(x_n^i, \langle x_n^{ij} \rangle_j), \langle f^{ij}(x_n^i, \langle x_n^{ijk} \rangle_k) \rangle_j \right) \\ &= \Phi_{f^i, \langle f^{ij} \rangle_j, \psi^i} \left(f^i \left(\Phi_{f^i, \langle f^{ij} \rangle_j, \psi^i}^{-1} (y_n^{i,(\kappa^i)}), \langle \Phi_{f^{ij}, \langle f^{ijk} \rangle_k, \psi^{ij}}^{-1} (y_n^{ij,(\kappa^{ij})}) \rangle_j \right) \right). \end{aligned} \quad (14)$$

Denote the RHS of (14) as $\mathbf{F}^i(y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j)$; the last $\kappa^i + \sum_j \kappa^{ij}$ components of \mathbf{F}^i are trivial. Denote the first component as $\mathbf{G}^i : \mathbb{R}^{\kappa^i} \times \mathbb{R}^{\sum_j \kappa^{ij}} \rightarrow \mathbb{R}$, then we arrive at (12). \square

Lemma 2. *Given an observed dataset $D = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$ where $\mathbf{y}_n \in \mathbb{R}^M$ are generated by a directed and acyclic synchronous GDS $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$, the 2TBN distribution can be written as*

$$\prod_{i=1}^M p_D(x_{n+1}^i \mid x_n^i, \langle x_n^{ij} \rangle_j) \cdot p_D(y_{n+1}^i \mid x_{n+1}^i) = \frac{\prod_{i=1}^M p_D(y_{n+1}^i \mid y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j)}{p_D(\mathbf{x}_n \mid \langle y_n^{i,(\kappa^i)} \rangle)}. \quad (15)$$

Proof. The generalised time delay embedding theorem [13] states that, under certain technical assumptions, and given M inhomogeneous observation functions $\{\psi^1, \psi^2, \dots, \psi^M\}$, the map

$$\Phi_{f,\psi}(\mathbf{x}) = \langle \Phi_{f^1, \psi^1}(\mathbf{x}), \Phi_{f^2, \psi^2}(\mathbf{x}), \dots, \Phi_{f^M, \psi^M}(\mathbf{x}) \rangle \quad (16)$$

is an embedding where each subsystem (local) map $\Phi_{f^i, \psi^i} : \mathcal{M} \rightarrow \mathbb{R}^{\kappa^i}$, and, at time index n is described by

$$\Phi_{f^i, \psi^i}(\mathbf{x}_n) = y_n^{i,(\kappa^i)} = \langle \psi^i(\mathbf{x}_n), \psi^i(\mathbf{x}_{n-\tau^i}), \psi^i(\mathbf{x}_{n-2\tau^i}), \dots, \psi^i(\mathbf{x}_{n-(\kappa^i-1)\tau^i}) \rangle$$

where $\sum_i \kappa^i = 2d + 1$ [13].² Therefore, the global map (16) is given by $\Phi_{f,\psi}(\mathbf{x}_n) = \langle y_n^{i,(\kappa^i)} \rangle$ and there must exist an inverse map $\mathbf{x}_n = \Phi_{f,\psi}^{-1}(\langle y_n^{i,(\kappa^i)} \rangle)$. Given Lemma 1, the existence of $\Phi_{f,\psi}^{-1}$, and

¹Stark [38] conjectures that the theorem should generalise to functions h that are not independent of x . To the best of our knowledge, this result remains to be proven.

²The original proof [13] uses positive lags, however the authors note that the use of negative lags also applies (and should be used in the case of endomorphisms [40]).

since $\forall i, \{y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j\} \subseteq \langle y_n^{i,(\kappa^i)} \rangle$, we arrive at the following equation:

$$\begin{aligned} \prod_{i=1}^M p_D \left(Y_{n+1}^i = \mathbf{G}^i \left(y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j \right) \mid y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j \right) \\ = p_D \left(\mathbf{X}_n = \Phi_{f,\psi}^{-1} \left(\langle y_n^{i,(\kappa^i)} \rangle \right) \mid \langle y_n^{i,(\kappa^i)} \rangle \right) \\ \times \prod_{i=1}^M p_D \left(X_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j) \mid x_n^i, \langle x_n^{ij} \rangle_j \right) \cdot p_D \left(Y_{n+1}^i = \psi^i(x_{n+1}^i) \mid x_{n+1}^i \right). \end{aligned} \quad (17)$$

Rearranging (17) gives the equality in (15). \square

Lemma 2 shows that the distributions can be reformulated by conditioning on delay vectors. The RHS of (15) can be used to perform inference in the 2TBN (7). The numerator is a product of local CPDs of scalar variables, and can thus be computed by either counting (for discrete variables) or density estimation (for continuous variables). The denominator is used to compute the probability that the hidden state occurred, given an observed delay vector; fortunately, Casdagli [41] established methods to compute this CPD for a variety of practical scenarios. Therefore, Lemma 2 provides a method to perform exact inference. Using this delay vector representation, we arrive at the following theorem.

Theorem 1. Consider a synchronous GDS $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$, where the graph G is a DAG. Each subsystem state follows the dynamics $x_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j)$ and generates an observation $y_{n+1}^i = \psi^i(x_{n+1}^i)$; a complete dataset is given by the sequence of observations $D = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$. The log-likelihood of the data given a network structure can be computed in terms of conditional entropy:

$$\ell(\hat{\Theta}_G : D) = N \cdot H(\mathbf{X}_n \mid \langle Y_n^{i,(\kappa^i)} \rangle) - N \cdot \sum_{i=1}^M H(Y_{n+1}^i \mid Y_n^{i,(\kappa^i)}, \langle Y_n^{ij,(\kappa^{ij})} \rangle_j) \quad (18)$$

Proof. Substituting (15) into (11) gives the log-likelihood $\ell(\hat{\Theta}_G : D)$ as

$$\begin{aligned} N \sum_{i=1}^M \sum_{y_{n+1}^i} \sum_{y_n^{i,(\kappa^i)}} \sum_{\langle y_n^{ij,(\kappa^{ij})} \rangle_j} p_D(y_{n+1}^i, y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j) \log p_D(y_{n+1}^i \mid y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j) \\ - N \sum_{\mathbf{x}_n} \sum_{\langle y_n^{i,(\kappa^i)} \rangle} p_D(\mathbf{x}_n, \langle y_n^{i,(\kappa^i)} \rangle) \log p_D(\mathbf{x}_n \mid \langle y_n^{i,(\kappa^i)} \rangle). \end{aligned} \quad (19)$$

In (19) we have removed arguments of the joint distributions that will be nullified when multiplied with the CPD. Expressing (19) in terms of conditional entropy (3), we arrive at (18). \square

Theorem 2. The information criterion (10) for synchronous GDS can be computed as:

$$\begin{aligned} g(B : D) = -N \cdot \sum_{i=1}^M H(Y_{n+1}^i \mid Y_n^{i,(\kappa^i)}, \langle Y_n^{ij,(\kappa^{ij})} \rangle_j) \\ - f(N) \cdot \sum_{i=1}^M \left(|Y_n^i|^{\kappa^i} \cdot (|Y_n^i| - 1) \cdot \prod_{V^p \in \Pi_G(V^i)} |Y_n^p|^{\kappa^p} \right). \end{aligned} \quad (20)$$

Proof. The distributions for the first term in (19) do not depend on the parents of a subsystem and thus are independent of the graph G being considered. Therefore, we have the following equation for maximum log-likelihood:

$$\max_G \ell(\hat{\Theta}_G : D) = \mathcal{O}(N) - N \cdot \min_G \sum_{i=1}^M H(Y_{n+1}^i \mid Y_n^{i,(\kappa^i)}, \langle Y_n^{ij,(\kappa^{ij})} \rangle_j). \quad (21)$$

We can now compute the number of parameters needed to specify the model as [28]

$$C(G) = \sum_{i=1}^M \left(|Y_n^i|^{\kappa^i} \cdot (|Y_n^i| - 1) \cdot \prod_{V^p \in \Pi_G(V^i)} |Y_n^p|^{\kappa^p} \right). \quad (22)$$

Since we are searching for the graph $G^* = \max_G g(B : D)$, holding N constant, we can substitute (21) and (22) into (10) and ignore the constant term $\mathcal{O}(N)$ in (21). \square

5.3 The log-likelihood and information transfer

To conclude our study of the scores, we look at the log-likelihood in the context of information transfer. First, rearranging the terms of collective transfer entropy (4) we can rewrite the log-likelihood function (18), leading to the following result.

Proposition 1. *The log-likelihood function for the synchronous GDS (18) decomposes as follows:*

$$\ell(\hat{\Theta}_G : D) = N \cdot H(\mathbf{X}_n | \langle Y_n^{i,(\kappa^i)} \rangle) - N \cdot \sum_{i=1}^M H(Y_{n+1}^i | Y_n^{i,(\kappa^i)}) + N \cdot \sum_{i=1}^M T_{\langle Y^{ij} \rangle_j \rightarrow Y^i}. \quad (23)$$

Again, the first two terms in (23) do not depend on the proposed graph structure, and thus maximising log-likelihood is equivalent to maximising collective transfer entropy. This becomes clear when we consider the *log-likelihood ratio*. This ratio quantifies the gain in likelihood by modelling the data D by a candidate network B instead of the empty network B_\emptyset , i.e.,

$$\ell(\hat{\Theta}_G : D) - \ell(\hat{\Theta}_{G_\emptyset} : D) \propto \log \frac{p(B | D)}{p(B_\emptyset | D)}.$$

Recall that the empty DAG G_\emptyset is one with no parents for all vertices $\forall i, \Pi_G(V^i) = \langle Y_n^{ij,(\kappa^{ij})} \rangle_j = \emptyset$. Substituting this definition into (18) (or, alternatively (23)) gives the following result.

Proposition 2. *The ratio of the log-likelihood (18) of a candidate DAG G to the empty network G_\emptyset can be expressed as*

$$\ell(\hat{\Theta}_G : D) - \ell(\hat{\Theta}_{G_\emptyset} : D) = N \cdot \sum_{i=1}^M T_{\langle Y^{ij} \rangle_j \rightarrow Y^i}.$$

6 Discussion and future work

We have presented a principled method to score the structure of nonlinear dynamical networks, where dynamical units are coupled via a DAG. We approached the problem by modelling the time evolution of a synchronous GDS as a DBN. We then derived the AIC and BIC scoring functions for the DBN based on time delay embedding theorems. Finally, we have shown that the log-likelihood of the synchronous GDS can be interpreted in the context of information transfer.

The representation of synchronous GDSs as DBNs allows for inference of coupling in dynamical networks and facilitates techniques for synthesis in these systems. DBNs are an expressive framework that allow representation of generic systems, as well as a numerous general purpose inference techniques that can be used for filtering, prediction, and smoothing [28]. Our representation therefore allows for probabilistic reasoning for purposes of planning and prediction in complex systems.

Theorem 2 captures an interesting parallel between learning from complete data and learning nonlinear dynamical networks. If the embedding dimension κ and time delay τ are unity, then the information criterion becomes identical to learning a DBN from complete data [28]. Thus, our result could be considered a generalisation of typical structure learning procedures.

The results presented here provoke new insights into the concepts of structure learning, nonlinear time series analysis and effective network analysis [17, 18] based on information transfer [19, 6, 20, 10]. The information-theoretic interpretation of the log-likelihood has interesting consequences in the context of information dynamics and information thermodynamics of nonlinear dynamical networks. The transfer entropy terms in Propositions 1 and 2 show that the optimal structure of a synchronous GDS is immediately related to the information processing of distributed computation [21], as well as the thermodynamic costs of information transfer [42].

In the future, we aim to perform empirical studies to exemplify the properties of the presented scoring functions. Specifically, the empirical studies should yield insight into the effect of weak, moderate and strong coupling between dynamical units. An important concept to consider in stochastic systems is the convergence of the shadow (reconstructed) manifold to the true manifold [33]; we have implicitly accounted for this phenomena by using CPDs in our model, however it is important to investigate the property of convergence with different density estimation techniques. In addition, we are interested in the effect of synchrony in these networks and the relationship to previous results for dynamical systems coupled by spanning trees [2]. We conjecture that approach used here will allow us to derive scoring functions without the assumption of multinomial observations, and thus afford the use of non-parametric density estimators. Parametric techniques, such as learning the parameters of dynamical systems [43, 44], could be considered in place of the posterior approximations.

Finally, the reconstruction theorems used in this paper typically make the assumption that the map (or flow) is a diffeomorphism (invertible in time). Thus, given any state, the past and future are uniquely determined and the time delay τ can be taken positive or negative. In certain cases, however, the time-reversed system is acausal, giving a map that is not time-invertible (an endomorphism). Ideally, we would aim to have methods to infer coupling for both endomorphisms and diffeomorphisms. Takens [40] showed that if the map is an endomorphism, taking the delay vector of temporally *previous* observations forms an embedding. The generalised theorems in [38, 12, 13], however, were established for diffeomorphisms, rather than endomorphisms; we can only conjecture that taking a delay of past observations (as we have done throughout this paper) follows for these results. Empirical studies using the measures presented in this paper would indicate whether it is an important line of inquiry to prove the generalised reconstruction theorems for endomorphisms.

Acknowledgements

We would like to thank Joseph Lizier, Jürgen Jost, and Wolfram Martens for many helpful discussions, particularly in regards to embedding theory.

This work was supported in part by the Australian Centre for Field Robotics; the New South Wales Government; and the Faculty of Engineering & Information Technologies, The University of Sydney, under the Faculty Research Cluster Program.

References

- [1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, “Complex networks: Structure and dynamics,” *Phys. Rep.*, vol. 424, no. 4, pp. 175–308, 2006.
- [2] C. W. Wu, “Synchronization in networks of nonlinear dynamical systems coupled via a directed graph,” *Nonlinearity*, vol. 18, no. 3, p. 1057, 2005.
- [3] H. Mortveit and C. Reidys, *An Introduction to Sequential Dynamical Systems*. Springer Science & Business Media, 2007.
- [4] H. Kantz and T. Schreiber, *Nonlinear time series analysis*. Cambridge university press, 2004.
- [5] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, “Transfer entropy – a model-free measure of effective connectivity for the neurosciences,” *J. Comp. Neurosci.*, vol. 30, no. 1, pp. 45–67, 2011.
- [6] J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, and M. Prokopenko, “Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity,” *J. Comp. Neurosci.*, vol. 30, no. 1, pp. 85–107, 2011.
- [7] J. Schumacher, T. Wunderle, P. Fries, F. Jäkel, and G. Pipa, “A statistical framework to infer delay and direction of information flow from measurements of complex systems,” *Neural Computation*, vol. 27, no. 8, pp. 1555–1608, 2015.
- [8] S. K. Gan, R. Fitch, and S. Sukkarieh, “Online decentralized information gathering with spatial–temporal constraints,” *Auton. Robots*, vol. 37, no. 1, pp. 1–25, 2014.
- [9] Z. Xu, R. Fitch, J. P. Underwood, and S. Sukkarieh, “Decentralized coordinated tracking with mixed discrete-continuous decisions,” *J. Field Robot.*, vol. 30, no. 5, pp. 717–740, 2013.

- [10] O. M. Cliff, J. T. Lizier, P. Wang, X. R. Wang, O. Obst, and M. Prokopenko, “Delayed spatio-temporal interactions and coherent structure in multi-agent team dynamics,” *Art. Life*, vol. 23, no. 1, 2016.
- [11] J. Umenberger and I. R. Manchester, “Scalable identification of stable positive systems,” in *Proc. of IEEE CDC*, 2016.
- [12] J. Stark, D. S. Broomhead, M. E. Davies, and J. Huke, “Delay embeddings for forced systems. II. Stochastic forcing,” *J. Nonlinear Sci.*, vol. 13, no. 6, pp. 519–577, 2003.
- [13] E. R. Deyle and G. Sugihara, “Generalized theorems for nonlinear state space reconstruction,” *PLOS ONE*, vol. 6, no. 3, p. e18295, 2011.
- [14] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
- [15] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [16] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, “Information modification and particle collisions in distributed computation,” *Chaos*, vol. 20, no. 3, pp. 037109–13, 2010.
- [17] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag, “Organization, development and function of complex brain networks,” *Trends Cogn. Sci.*, vol. 8, no. 9, pp. 418–425, 2004.
- [18] H.-J. Park and K. Friston, “Structural and functional brain networks: from connections to cognition,” *Science*, vol. 342, no. 6158, p. 1238411, 2013.
- [19] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns, “Network structure of cerebral cortex shapes functional connectivity on multiple time scales,” *Proc. of Natl. Acad. Sci.*, vol. 104, no. 24, pp. 10240–10245, 2007.
- [20] O. M. Cliff, J. T. Lizier, X. R. Wang, P. Wang, O. Obst, and M. Prokopenko, “Towards quantifying interaction networks in a football match,” in *RoboCup 2013: Robot World Cup XVII*, pp. 1–13, Springer, 2013.
- [21] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, “Local information transfer as a spatiotemporal filter for complex systems,” *Phys. Rev. E*, vol. 77, no. 2, p. 026110, 2008.
- [22] L. Kocarev and U. Parlitz, “Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems,” *Physical Review Letters*, vol. 76, no. 11, p. 1816, 1996.
- [23] K. Kaneko, “Overview of coupled map lattices,” *Chaos*, vol. 2, no. 3, pp. 279–282, 1992.
- [24] D. M. Chickering, “Learning equivalence classes of Bayesian-network structures,” *J. Mach. Learn. Res.*, vol. 2, pp. 445–498, 2002.
- [25] W. Lam and F. Bacchus, “Learning Bayesian belief networks: An approach based on the MDL principle,” *Comp. Intell.*, vol. 10, no. 3, pp. 269–293, 1994.
- [26] R. R. Bouckaert, “Properties of Bayesian belief network learning algorithms,” in *Proc. of AUAI UAI*, pp. 102–109, 1994.
- [27] D. Heckerman, D. Geiger, and D. M. Chickering, “Learning Bayesian networks: the combination of knowledge and statistical data,” *Mach. Learn.*, vol. 20, no. 3, pp. 20–197, 1995.
- [28] N. Friedman, K. Murphy, and S. Russell, “Learning the structure of dynamic probabilistic networks,” in *Proc. of AUAI UAI*, pp. 139–147, 1998.
- [29] Z. Ghahramani, “Learning dynamic Bayesian networks,” in *Adaptive Processing of Sequences and Data Structures*, vol. 1387 of *Lecture Notes in Comp. Sci.*, pp. 168–197, 1998.
- [30] T. Schreiber, “Measuring information transfer,” *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461–464, 2000.
- [31] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, “Nonlinear causal discovery with additive noise models,” in *Advances in Neural Information Processing Systems 21*, pp. 689–696, Curran Associates, Inc., 2009.
- [32] C. W. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica*, pp. 424–438, 1969.

- [33] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, “Detecting causality in complex ecosystems,” *Science*, vol. 338, no. 6106, pp. 496–500, 2012.
- [34] J. T. Lizier and M. Rubinov, “Multivariate construction of effective computational networks from observational data.” ArXiv Preprint, 2012.
- [35] J. Peters, D. Janzing, and B. Schölkopf, “Causal inference on discrete data using additive noise models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2436–2450, 2011.
- [36] A. Gretton, P. Spirtes, and R. E. Tillman, “Nonlinear directed acyclic structure learning with weakly additive noise models,” in *Advances in Neural Information Processing Systems 22*, pp. 1847–1855, Curran Associates, Inc., 2009.
- [37] F. Takens, “Detecting strange attractors in turbulence,” in *Dynamical Systems and Turbulence*, vol. 898 of *Lecture Notes in Math.*, pp. 366–381, 1981.
- [38] J. Stark, “Delay embeddings for forced systems. I. Deterministic forcing,” *J. Nonlinear Sci.*, vol. 9, no. 3, pp. 255–332, 1999.
- [39] H. S. Mortveit and C. M. Reidys, “Discrete, sequential dynamical systems,” *Discrete Math.*, vol. 226, no. 1, pp. 281–295, 2001.
- [40] F. Takens, “The reconstruction theorem for endomorphisms,” *Bull. Braz. Math. Soc.*, vol. 33, no. 2, pp. 231–262, 2002.
- [41] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson, “State space reconstruction in the presence of noise,” *Physica D.*, vol. 51, no. 1, pp. 52–98, 1991.
- [42] M. Prokopenko and J. T. Lizier, “Transfer entropy and transient limits of computation,” *Sci. Rep.*, vol. 4, p. 5394, 2014.
- [43] Z. Ghahramani and S. T. Roweis, “Learning nonlinear dynamical systems using an EM algorithm,” in *Advances in Neural Information Processing Systems 11*, pp. 431–437, MIT Press, 1999.
- [44] A. Hefny, C. Downey, and G. J. Gordon, “Supervised learning for dynamical system learning,” in *Advances in Neural Information Processing Systems 28*, pp. 1963–1971, Curran Associates, Inc., 2015.